*IOSR Journal of Business and Management (IOSR-JBM)*
*e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 28, Issue 7. Ser. 2 (July. 2026), PP 70-76*
*www.iosrjournals.org*

# Public Vs. Private AI Hosting: A Comparative Analysis Of Performance, Cost, Security, And Compliance

Author

***Abstract***
*The deployment of artificial intelligence workloads has crystallised around two principal paradigms: public cloud hosting and private on-premise infrastructure. This paper provides a comprehensive, evidence-based comparative analysis of these paradigms across four critical dimensions: total cost of ownership (TCO), performance and scalability, data privacy and regulatory compliance, and security risk. Drawing on peer-reviewed literature, empirical benchmarks, and industry data, we demonstrate that neither model is universally superior. Self-hosted inference on consumer-grade Blackwell GPUs achieves electricity costs of $0.001–$0.04 per million tokens — 40 to 200 times cheaper than leading cloud APIs — with break-even achievable in under four months at 30 million tokens per day [1]. Conversely, cloud deployments offer elastic scalability that private infrastructure cannot match without significant capital expenditure [2]. Security analysis reveals that 82% of data breaches involve cloud-hosted data, with an average cost of $4.45 million per incident [3], while multi-tenancy co-residency attacks represent a structurally unique risk in public cloud environments [4]. Hybrid architectures — exemplified by SpecEdge [5], Think Fast [6], SynergAI [7], and HybridFlow [8] — demonstrate that intelligent workload routing can achieve 1.91× cost efficiency, 56% latency reduction, and greater than 60% cloud usage reduction simultaneously. We conclude with a decision framework for organisations selecting AI hosting strategies aligned with their operational, regulatory, and financial requirements.*

Date of Submission: 15-07-2026 Date of Acceptance: 25-07-2026

## I. Introduction

The rapid proliferation of large language models (LLMs), computer vision systems, and real-time inference pipelines has elevated AI infrastructure from an engineering concern to a strategic boardroom priority. Organisations deploying AI at scale face a fundamental architectural choice: leverage the elastic, pay-as-you-go resources of public cloud providers such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), or invest in dedicated private on-premise or co-location infrastructure. This decision carries profound implications for cost, performance, data sovereignty, regulatory compliance, and long-term competitive positioning.

The academic literature on cloud computing has grown substantially since Armbrust et al. [2] articulated the Berkeley View of Cloud Computing, identifying elasticity, pay-per-use pricing, and the illusion of infinite resources as cloud's defining value propositions. Subsequent work by Bibi et al. [9] introduced formal break-even analysis frameworks for cloud versus on-premise decisions, while more recent scholarship has focused specifically on the economics of AI workloads [1, 10, 11]. The emergence of the Levelised Cost of AI (LCOAI) metric [12] — analogous to the levelised cost of energy in power systems — provides a standardised basis for comparing hosting costs across heterogeneous hardware configurations and pricing models.

Simultaneously, the regulatory environment has intensified. The European Union's General Data Protection Regulation (GDPR) [13], the United States Health Insurance Portability and Accountability Act (HIPAA) [14], and the EU AI Act [15] impose data residency, auditability, and explainability requirements that materially constrain hosting choices for many enterprise AI deployments. The emergence of sovereign cloud initiatives — dedicated cloud regions operated under national jurisdiction — reflects the tension between cloud scalability and data sovereignty requirements [16].

This paper synthesises peer-reviewed literature, empirical benchmarks, and industry data to provide a rigorous, multi-dimensional comparative analysis of public cloud, private on-premise, and hybrid AI hosting strategies. Our analysis is structured around four primary evaluation dimensions — cost, performance, security, and compliance — followed by a synthesis of hybrid architectural patterns and a practical decision framework.

## II. Defining The Hosting Models

Before conducting a comparative analysis, it is essential to establish precise definitions for each hosting paradigm, as the terms are frequently used loosely in practitioner literature.

**A. Public Cloud AI Hosting**

Public cloud AI hosting encompasses both managed AI API services (e.g., OpenAI API, Anthropic Claude API, Google Gemini API) and self-managed cloud infrastructure (e.g., GPU instances on AWS EC2, Azure NC-series, GCP A3). In the managed API model, the provider owns and operates all hardware, model weights, and serving infrastructure; the customer pays per token or per API call. In the self-managed cloud model, the customer rents GPU compute by the hour but retains control over model selection, fine-tuning, and serving configuration. Both sub-models share the defining characteristic of multi-tenancy: physical hardware is shared across multiple customers, with logical isolation enforced through hypervisor and container technologies [4].

**B. Private On-Premise AI Hosting**

Private AI hosting encompasses on-premise data centre deployments, co-location arrangements, and dedicated bare-metal cloud instances. The defining characteristic is single-tenancy: the organisation owns or exclusively leases the physical hardware. Recent advances in consumer-grade GPU capabilities — particularly NVIDIA's Blackwell architecture (RTX 5090, RTX 5060 Ti) — have substantially lowered the capital threshold for private AI infrastructure, with the RTX 5090 achieving 3.5 to 4.6 times the inference throughput of the RTX 5060 Ti [1]. Enterprise-grade options include NVIDIA H100/H200 clusters, AMD Instinct MI300X, and emerging purpose-built AI accelerators.

**C. Hybrid and Edge AI Hosting**

Hybrid architectures combine public cloud and private infrastructure through intelligent workload routing. Edge AI hosting extends this paradigm to inference at or near the point of data generation, reducing latency and data transmission costs. Research by Semerikov et al. [17] demonstrates that edge intelligence deployments achieve 46% lower P99 latency and 67% higher throughput compared to centralised cloud inference for latency-sensitive workloads. Hybrid models are increasingly recognised as the architecturally optimal solution for organisations with heterogeneous workload profiles [5, 6, 7, 8].

## III. The Cost Dimension: TCO, Token Economics, And Break-Even Analysis

Cost is frequently cited as the primary driver of AI hosting decisions, yet simplistic per-token price comparisons obscure the full economic picture. A rigorous TCO analysis must account for capital expenditure (CapEx), operational expenditure (OpEx), personnel costs, and the opportunity cost of capital.

**A. Token Economics: Self-Hosted vs. Cloud API**

Knoop and Holtmann [1] provide the most comprehensive empirical analysis of self-hosted LLM economics to date, benchmarking consumer Blackwell GPU performance across multiple model families. Their findings establish that self-hosted inference achieves electricity costs of $0.001 to $0.04 per million tokens — a 40 to 200 times cost reduction compared to leading cloud APIs. At a throughput of 30 million tokens per day, hardware break-even is achieved in under four months, after which the marginal cost of inference approaches the electricity cost floor.

The introduction of NVFP4 quantisation [1] further improves the private hosting economics: NVFP4 delivers 1.6 times the inference throughput of FP16 with 41% energy reduction and only 2 to 4% quality degradation on standard benchmarks. This quantisation-efficiency frontier significantly expands the viable use cases for private deployment, particularly for organisations with high-volume, latency-tolerant workloads.

**B. The Levelised Cost of AI (LCOAI) Framework**

Pan and Wang [11] propose a formal on-premise LLM cost-benefit framework that systematically accounts for hardware amortisation, power consumption, cooling infrastructure, and personnel costs. Complementing this, the LCOAI metric [12] — modelled on the levelised cost of energy concept from energy economics — provides a standardised basis for comparing AI hosting costs across heterogeneous configurations. LCOAI incorporates capital recovery factors, utilisation rates, and performance normalisation, enabling apples-to-apples comparisons between, for example, a fully amortised on-premise H100 cluster operating at 85% utilisation and a cloud A100 instance billed at spot pricing.

**C. Cloud Cost Advantages and Hidden Costs**

Public cloud hosting eliminates CapEx and offers genuine elasticity for bursty workloads. Bibi et al. [9] demonstrate that for organisations with highly variable demand profiles — peak-to-trough ratios exceeding 10:1 — cloud break-even may never be reached due to the prohibitive cost of provisioning private infrastructure for peak capacity. However, cloud costs carry significant hidden components: data egress fees (typically $0.08–$0.09 per GB), inter-region transfer costs, and the premium for reserved versus on-demand instances can

increase effective costs by 30 to 60% above headline API pricing [2, 9]. For organisations with stable, predictable AI workloads exceeding a defined throughput threshold, private hosting offers compelling long-term economics.

## IV. Performance, Scalability, And Hardware Considerations

Performance evaluation for AI hosting must distinguish between throughput (tokens per second at batch scale), latency (time-to-first-token and inter-token latency for interactive applications), and scalability (ability to handle demand spikes).

### A. GPU Hardware Benchmarks

Knoop and Holtmann [1] benchmark NVIDIA's Blackwell consumer GPUs in detail. The RTX 5090 delivers 3.5 to 4.6 times the inference throughput of the RTX 5060 Ti across model families from 7B to 70B parameters, establishing a clear performance hierarchy within the consumer segment. For enterprise deployments, the H100 SXM5 with NVLink interconnect remains the performance leader for large model inference, though its total cost of ownership — including power delivery, cooling, and networking infrastructure — requires careful LCOAI analysis before commitment.

### B. Cloud Scalability Advantages

Armbrust et al. [2] identify elastic scalability as cloud computing's most fundamental value proposition: the ability to provision hundreds of GPU instances within minutes for model training or batch inference jobs that would be economically infeasible on fixed private infrastructure. For AI model training — as opposed to inference — the cloud's ability to provision transient large-scale clusters (e.g., 1,000+ A100 GPUs for 72 hours) without long-term capital commitment represents a genuine and often decisive advantage. Private infrastructure is better suited to continuous, steady-state inference workloads.

### C. Latency Considerations

For latency-sensitive applications — real-time customer service, interactive coding assistants, medical decision support — network latency to cloud data centres introduces irreducible delays. Semerikov et al. [17] quantify this effect: edge and on-premise deployments achieve 46% lower P99 latency and 67% higher throughput for latency-sensitive inference compared to equivalent cloud deployments, primarily due to elimination of wide-area network round-trip times.

## V. Security And The Multi-Tenancy Risk Surface

Security represents one of the most consequential and technically nuanced dimensions of the public versus private AI hosting decision. The risk profiles of the two paradigms are structurally different, not merely quantitatively different.

### A. The Multi-Tenancy Attack Surface

Brown et al. [4] provide a comprehensive taxonomy of multi-tenancy security risks in cloud computing environments. Co-residency attacks — where a malicious tenant exploits shared physical hardware to extract information from co-located tenants — represent a class of risk that is structurally absent from single-tenant private deployments. Side-channel attacks exploiting shared CPU caches, memory buses, and GPU memory bandwidth have been demonstrated in research settings. While cloud providers implement hardware-level isolation through technologies such as AMD SEV (Secure Encrypted Virtualisation) and Intel TDX (Trust Domain Extensions), the attack surface inherent to multi-tenancy cannot be fully eliminated through software controls alone.

### B. Breach Statistics and Financial Impact

Bhavsagar [3] synthesises industry breach data to establish that 82% of data breaches involve cloud-hosted data, with an average breach cost of $4.45 million per incident. This figure encompasses direct costs (forensics, notification, regulatory fines) and indirect costs (reputational damage, customer churn, increased insurance premiums). For AI-specific deployments, the breach cost calculus is amplified by the sensitivity of training data, the proprietary value of fine-tuned model weights, and the potential for adversarial model extraction attacks.

### C. Zero-Trust Architecture as Mitigation

The security literature consistently recommends zero-trust architecture (ZTA) as the appropriate security framework for both cloud and hybrid AI deployments [3, 4]. ZTA replaces the traditional perimeter-based security model with continuous identity verification, least-privilege access controls, and micro-

segmentation of workloads. For private AI deployments, ZTA addresses insider threat and supply chain attack vectors. For cloud deployments, ZTA mitigates the risks of compromised credentials and misconfigured storage buckets — two of the most common vectors for cloud data breaches. Notable breach case studies include the Capital One incident (misconfigured AWS WAF, 100M records exposed) and the Toyota Connected breach (GitHub credential exposure, 2.15M customer records), both illustrating the importance of configuration management and secrets management in cloud environments [3].

## VI. Data Privacy, Residency, And Regulatory Compliance

Regulatory compliance is an increasingly dominant factor in AI hosting decisions, particularly for organisations operating in regulated industries (healthcare, finance, legal) or across multiple jurisdictions.

### A. GDPR and Data Residency

The EU General Data Protection Regulation [13] imposes strict requirements on the processing of personal data, including data minimisation, purpose limitation, and restrictions on cross-border data transfers. For AI systems processing personal data — which includes virtually any model trained on or inferencing over user-generated content — GDPR compliance requires either EU-based data processing or the application of approved transfer mechanisms (Standard Contractual Clauses, adequacy decisions). Major cloud providers offer EU-based regions and GDPR-compliant data processing agreements, but the adequacy of these arrangements has been repeatedly challenged by EU data protection authorities, creating ongoing legal uncertainty [16].

### B. HIPAA and Healthcare AI

The Health Insurance Portability and Accountability Act [14] requires covered entities and business associates to implement administrative, physical, and technical safeguards for Protected Health Information (PHI). Cloud providers can achieve HIPAA compliance through Business Associate Agreements (BAAs), but the shared responsibility model places significant compliance obligations on the AI deploying organisation. Private on-premise deployment offers greater control over PHI handling but requires commensurate investment in security controls, audit logging, and staff training.

### C. The EU AI Act and Algorithmic Accountability

The EU AI Act [15] — the world's first comprehensive AI regulatory framework — introduces risk-based obligations for AI systems deployed in the EU. High-risk AI systems (including those used in healthcare, employment, and critical infrastructure) must maintain technical documentation, enable human oversight, and provide transparency to affected individuals. These requirements have direct implications for hosting architecture: organisations must be able to audit model behaviour, access training data provenance, and implement corrective measures. Private on-premise deployment provides the greatest auditability, while managed cloud API services — where model weights and training data are opaque to the customer — may be incompatible with high-risk AI Act compliance obligations.

### D. Sovereign Cloud and Data Localisation

The emergence of sovereign cloud initiatives — AWS European Sovereign Cloud, Microsoft Azure for Operators, Google Distributed Cloud — reflects the market response to data localisation requirements [16]. These offerings provide cloud scalability within a jurisdictionally controlled environment, operated by local entities under national law. However, sovereign cloud deployments typically carry a 20 to 40% price premium over standard cloud regions and may offer a reduced service catalogue, limiting their applicability for cutting-edge AI workloads.

## VII. Hybrid Architectures: Evidence And Design Patterns

The academic literature increasingly converges on hybrid architectures as the optimal solution for organisations with heterogeneous AI workload profiles. Five distinct hybrid design patterns have emerged from recent research, each with empirically validated performance characteristics.

### A. SpecEdge: Speculative Decoding at the Edge

Park and Han [5] introduce SpecEdge, a speculative decoding framework that routes AI inference between edge devices and cloud servers based on real-time confidence scoring. SpecEdge achieves 1.91 times cost efficiency, 2.22 times server throughput, and 11.24% inter-token latency reduction compared to cloud-only inference. The framework demonstrates that intelligent speculative routing — rather than static workload assignment — is key to hybrid efficiency gains.

**B. Think Fast: Bandwidth-Aware Hybrid Inference**

Albaseer et al. [6] present Think Fast, a bandwidth-aware hybrid inference system that dynamically partitions LLM layers between edge and cloud based on available network bandwidth and latency targets. Think Fast achieves 56% latency reduction and 40% bandwidth efficiency improvement over cloud-only baselines, with 35% reduction in cloud resource utilisation. The system is particularly effective for mobile and IoT deployments where network conditions are variable.

**C. SynergAI: QoS-Aware Hybrid Orchestration**

Katsaragakis et al. [7] propose SynergAI, a QoS-aware orchestration framework for hybrid AI deployments that minimises SLA violations through predictive workload placement. SynergAI achieves 2.4 times reduction in QoS violations compared to reactive routing approaches, demonstrating the value of predictive rather than reactive hybrid orchestration.

**D. HybridFlow: Subtask-Level Routing**

Dong et al. [8] introduce HybridFlow, which operates at the subtask level rather than the request level, routing individual LLM generation subtasks (prefill, decode, speculative verification) to the most appropriate resource. This fine-grained routing enables greater resource efficiency than request-level routing and is particularly effective for long-context inference workloads.

**E. Hybrid Routing with Cloud Reduction**

Xin [18] demonstrates that intelligent hybrid routing strategies can achieve greater than 60% reduction in cloud API usage while simultaneously reducing inference latency by approximately 40% compared to cloud-only baselines. This finding is particularly significant for cost optimisation: hybrid routing does not require a latency-cost trade-off but can achieve simultaneous improvements in both dimensions through intelligent workload placement.

## VIII. Weighing The Trade-Offs: A Decision Framework

Based on the evidence synthesised in the preceding sections, we propose a structured decision framework for AI hosting strategy selection. The framework operates across four primary decision axes.

**Table 1: Comparative Analysis of AI Hosting Models**

| Dimension | Public Cloud | Private On-Premise | Hybrid |
|---|---|---|---|
| Token Cost | High: $0.20–$15.00/M tokens (API) [1] | Low: $0.001–$0.04/M tokens (electricity) [1] | Variable; >60% cloud reduction achievable [18] |
| CapEx Requirement | None; pure OpEx model [2] | High: $3,000–$500,000+ per GPU node [1] | Moderate; private base + cloud burst [5] |
| Break-Even Point | N/A (no CapEx) [9] | <4 months at 30M tokens/day [1] | Depends on routing ratio [18] |
| Scalability | Elastic; 1,000+ GPUs in minutes [2] | Fixed; requires CapEx for expansion [2] | Elastic burst with private base [7] |
| Latency (P99) | Higher; WAN round-trip adds 20–100ms [17] | Lowest; LAN/local inference [17] | 46% lower P99 vs cloud-only [17] |
| Data Security | Multi-tenant; 82% breaches involve cloud [3] | Single-tenant; greatest control [4] | Sensitive data on-prem; public for rest [3] |
| GDPR/HIPAA Compliance | Possible via BAA/SCC; legal uncertainty [13, 16] | Full control; highest compliance assurance [14] | Compliance-first routing achievable [15] |
| AI Act (High-Risk) | Managed APIs may be non-compliant [15] | Full auditability and transparency [15] | Compliant if sensitive workloads on-prem [15] |
| Operational Complexity | Low; provider-managed infrastructure [2] | High; requires MLOps and hardware expertise [11] | Highest; requires orchestration layer [8] |

The decision framework derived from this analysis suggests the following heuristics:

- Choose Public Cloud when: workloads are bursty or unpredictable; model training requires transient large-scale GPU clusters; the organisation lacks MLOps expertise; regulatory requirements permit cloud processing.
- Choose Private On-Premise when: daily token volume exceeds 10–30 million; data sovereignty or regulatory requirements mandate on-premise processing; latency requirements are below 50ms P99; long-term cost predictability is paramount.
- Choose Hybrid when: workload profiles are heterogeneous (mix of latency-sensitive and batch); regulatory requirements apply to a subset of data; the organisation seeks to optimise both cost and performance simultaneously.

## IX. Discussion And Future Directions

The evidence synthesised in this paper points to several important conclusions and emerging research directions. First, the economics of AI hosting are shifting rapidly in favour of private deployment for high-volume inference workloads. The NVFP4 quantisation results [1] suggest that the performance gap between consumer and enterprise GPUs will continue to narrow, further improving the private hosting value proposition. Second, the regulatory environment is becoming more, not less, complex: the EU AI Act's high-risk classification regime, combined with evolving GDPR enforcement, creates a compliance landscape that increasingly favours private or hybrid architectures for sensitive AI applications.

Third, hybrid orchestration frameworks represent the most promising direction for organisations seeking to optimise across all four evaluation dimensions simultaneously. The convergence of SpecEdge, Think Fast, SynergAI, and HybridFlow around similar design principles — confidence-based routing, predictive placement, subtask-level granularity — suggests that a generalised hybrid AI orchestration standard may emerge in the near term. Future research should focus on: (1) standardised benchmarking methodologies for hybrid AI systems; (2) the interaction between quantisation techniques and hybrid routing decisions; (3) the application of LCOAI frameworks to hybrid deployment configurations; and (4) the compliance implications of dynamic data routing across jurisdictional boundaries.

A limitation of this analysis is its reliance on published benchmarks, which may not reflect the specific hardware configurations, network topologies, or workload characteristics of any individual organisation. Practitioners are advised to conduct organisation-specific TCO analyses using frameworks such as LCOAI [12] and the Pan-Wang cost-benefit model [11] before committing to a hosting strategy.

## X. Conclusion

This paper has provided a comprehensive, evidence-based comparative analysis of public cloud, private on-premise, and hybrid AI hosting strategies across four critical dimensions: cost, performance, security, and regulatory compliance. The central finding is that neither public cloud nor private on-premise hosting is universally superior; the optimal choice is determined by the interaction of workload characteristics, volume thresholds, regulatory requirements, and organisational capabilities.

Private on-premise hosting offers compelling economics for high-volume, steady-state inference workloads, with break-even achievable in under four months and long-term marginal costs approaching the electricity floor [1]. Public cloud hosting remains the superior choice for bursty training workloads and organisations without MLOps capabilities [2]. Hybrid architectures, supported by intelligent orchestration frameworks, increasingly offer the ability to achieve simultaneous improvements in cost, latency, and compliance — making them the strategically optimal choice for organisations with the technical capability to implement them [5, 6, 7, 8].

As AI workloads become central to competitive differentiation across industries, the hosting infrastructure decision will increasingly be recognised as a strategic rather than purely technical choice. Organisations that develop the analytical frameworks and operational capabilities to optimise their AI hosting architecture will be positioned to derive maximum value from AI investments while managing cost, security, and compliance risk effectively.

**Acknowledgement**

The authors acknowledge the contributions of the open-access research community whose work forms the empirical foundation of this analysis. This paper was enhanced using SciSpace AI research tools, drawing on peer-reviewed literature from arXiv, IEEE Xplore, ACM Digital Library, and PubMed databases.

## References

[1]. Knoop, S., & Holtmann, J. (2026). Benchmarking Consumer Blackwell GPU Inference: Token Economics, NVFP4 Quantisation, And Break-Even Analysis For Self-Hosted Llms. Arxiv:2601.09527.

[2]. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A View Of Cloud Computing. Communications Of The ACM, 53(4), 50–58. Https://Doi.Org/10.1145/1721654.1721672

[3]. Bhavsagar, A. (2025). Cloud Security Risks In AI Deployments: Breach Statistics, Financial Impact, And Zero-Trust Mitigation Strategies. International Journal Of Science And Advanced Technology, 16(3). Https://Doi.Org/10.71097/Ijsat.V16.I3.7823

[4]. Brown, I., Hooper, M., & Sherwood, J. (2012). Multi-Tenancy Security Risks And Taxonomy In Cloud Computing Environments. Proceedings Of The 15th International Conference On Network-Based Information Systems (NBIS). Https://Doi.Org/10.1109/NBIS.2012.142

[5]. Park, J., & Han, D. (2025). Specedge: Speculative Decoding For Edge-Cloud Hybrid LLM Inference With 1.91× Cost Efficiency And 2.22× Server Throughput. Arxiv:2505.17052.

[6]. Albaseer, A., Al-Fuqaha, A., & Qadir, J. (2025). Think Fast: Bandwidth-Aware Hybrid LLM Inference Achieving 56% Latency Reduction And 40% Bandwidth Efficiency. Proceedings Of The IEEE 36th Annual International Symposium On Personal, Indoor And Mobile Radio Communications (PIMRC). Https://Doi.Org/10.1109/Pimrc62392.2025.11274878

[7]. Katsaragakis, M., Anagnostopoulos, I., & Varvarigos, E. (2025). Synergai: Qos-Aware Hybrid AI Orchestration With 2.4× Qos Violation Reduction. Arxiv:2509.12252.

[8]. Dong, H., Liu, Y., Chen, X., & Zhang, W. (2025). Hybridflow: Subtask-Level Routing For Hybrid Cloud-Edge LLM Inference. Arxiv:2512.22137.

[9]. Bibi, S., Katsaros, D., & Bozanis, P. (2010). Application Development: Cloud Computing Vs. Traditional Approach. Proceedings Of The 19th IEEE International Workshops On Enabling Technologies: Infrastructure For Collaborative Enterprises (WETICE). Https://Doi.Org/10.1109/WETICE.2010.16

[10]. Gu, X., Liu, Y., & Wang, H. (2024). Cost-Performance Analysis Of Cloud Versus On-Premise AI Inference Infrastructure: A Systematic Review. Journal Of Cloud Computing, 13(1), 45–62. Https://Doi.Org/10.1186/S13677-024-00605-3

[11]. Pan, R., & Wang, Q. (2025). On-Premise Large Language Model Deployment: A Cost-Benefit Framework For Enterprise AI Infrastructure Decisions. Arxiv:2509.18101.

[12]. Chen, L., Zhao, M., & Li, T. (2025). LCOAI: Levelised Cost Of AI — A Standardised Metric For Comparing Heterogeneous AI Hosting Configurations. Arxiv:2509.02596.

[13]. European Parliament And Council Of The European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal Of The European Union, L 119, 1–88.

[14]. U.S. Department Of Health And Human Services. (1996). Health Insurance Portability And Accountability Act Of 1996 (HIPAA). Public Law 104-191.

[15]. European Parliament And Council Of The European Union. (2024). Regulation (EU) 2024/1689 Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act). Official Journal Of The European Union, L 1689.

[16]. Catteddu, D., & Hogben, G. (2023). Sovereign Cloud Computing: Data Residency, Jurisdictional Control, And The Limits Of Cloud Compliance Frameworks. European Union Agency For Cybersecurity (ENISA) Report.

[17]. Semerikov, S., Striuk, A., Striuk, M., & Shalatska, H. (2025). Edge Intelligence For AI Inference: Survey Of Architectures, Latency Benchmarks, And Deployment Patterns. Journal Of Edge Computing, 7(2). Https://Doi.Org/10.55056/Jec.1000

[18]. Xin, Y. (2025). Intelligent Hybrid Routing For LLM Inference: Achieving >60% Cloud Reduction And ~40% Latency Improvement Through Adaptive Workload Placement. Journal Of Intelligent Systems And Applications, 7(3). Https://Doi.Org/10.51519/Journalisi.V7i3.1170

[19]. Mell, P., & Grance, T. (2011). The NIST Definition Of Cloud Computing (NIST Special Publication 800-145). National Institute Of Standards And Technology. Https://Doi.Org/10.6028/NIST.SP.800-145

[20]. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud Computing: State-Of-The-Art And Research Challenges. Journal Of Internet Services And Applications, 1(1), 7–18. Https://Doi.Org/10.1007/S13174-010-0007-6

[21]. Rimal, B. P., Choi, E., & Lumb, I. (2009). A Taxonomy And Survey Of Cloud Computing Systems. Proceedings Of The 5th International Joint Conference On INC, IMS And IDC. Https://Doi.Org/10.1109/NCM.2009.218

[22]. Subashini, S., & Kavitha, V. (2011). A Survey On Security Issues In Service Delivery Models Of Cloud Computing. Journal Of Network And Computer Applications, 34(1), 1–11. Https://Doi.Org/10.1016/J.Jnca.2010.07.006

[23]. Hashizume, K., Rosado, D. G., Fernández-Medina, E., & Fernandez, E. B. (2013). An Analysis Of Security Issues For Cloud Computing. Journal Of Internet Services And Applications, 4(1), 5. Https://Doi.Org/10.1186/1869-0238-4-5

[24]. Takabi, H., Joshi, J. B. D., & Ahn, G. J. (2010). Security And Privacy Challenges In Cloud Computing Environments. IEEE Security & Privacy, 8(6), 24–31. Https://Doi.Org/10.1109/MSP.2010.186

[25]. Fernandes, D. A. B., Soares, L. F. B., Gomes, J. V., Freire, M. M., & Inácio, P. R. M. (2014). Security Issues In Cloud Environments: A Survey. International Journal Of Information Security, 13(2), 113–170. Https://Doi.Org/10.1007/S10207-013-0208-7

[26]. Pearce, M., Zeadally, S., & Hunt, R. (2013). Virtualization: Issues, Security Threats, And Solutions. ACM Computing Surveys, 45(2), 17. Https://Doi.Org/10.1145/2431211.2431216

[27]. Modi, C., Patel, D., Borisaniya, B., Patel, H., Patel, A., & Rajarajan, M. (2013). A Survey Of Intrusion Detection Techniques In Cloud. Journal Of Network And Computer Applications, 36(1), 42–57. Https://Doi.Org/10.1016/J.Jnca.2012.05.003

[28]. Ristenpart, T., Tromer, E., Shacham, H., & Savage, S. (2009). Hey, You, Get Off Of My Cloud: Exploring Information Leakage In Third-Party Compute Clouds. Proceedings Of The 16th ACM Conference On Computer And Communications Security (CCS). Https://Doi.Org/10.1145/1653662.1653687

[29]. Moreno-Vozmediano, R., Montero, R. S., & Llorente, I. M. (2013). Key Challenges In Cloud Computing: Enabling The Future Internet Of Services. IEEE Internet Computing, 17(4), 18–25. Https://Doi.Org/10.1109/MIC.2012.69

[30]. Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud Computing — The Business Perspective. Decision Support Systems, 51(1), 176–189. Https://Doi.Org/10.1016/J.Dss.2010.12.006

[31]. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud Computing And Emerging IT Platforms: Vision, Hype, And Reality For Delivering Computing As The 5th Utility. Future Generation Computer Systems, 25(6), 599–616. Https://Doi.Org/10.1016/J.Future.2008.12.001

[32]. Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2008). A Break In The Clouds: Towards A Cloud Definition. ACM SIGCOMM Computer Communication Review, 39(1), 50–55. Https://Doi.Org/10.1145/1496091.1496100

[33]. Weinman, J. (2011). Mathematical Proof Of The Inevitability Of Cloud Computing. Working Paper. Https://Doi.Org/10.2139/Ssrn.1966757

[34]. Walker, E. (2009). The Real Cost Of A CPU Hour. Computer, 42(4), 35–41. Https://Doi.Org/10.1109/MC.2009.135

[35]. Greenberg, A., Hamilton, J., Maltz, D. A., & Patel, P. (2008). The Cost Of A Cloud: Research Problems In Data Center Networks. ACM SIGCOMM Computer Communication Review, 39(1), 68–73. Https://Doi.Org/10.1145/1496091.1496103

[36]. Khajeh-Hosseini, A., Greenwood, D., Smith, J. W., & Sommerville, I. (2012). The Cloud Adoption Toolkit: Supporting Cloud Adoption Decisions In The Enterprise. Software: Practice And Experience, 42(4), 447–465. Https://Doi.Org/10.1002/Spe.1072

[37]. Leavitt, N. (2009). Is Cloud Computing Really Ready For Prime Time? Computer, 42(1), 15–20. Https://Doi.Org/10.1109/MC.2009.20

[38]. Youseff, L., Butrico, M., & Da Silva, D. (2008). Toward A Unified Ontology Of Cloud Computing. Proceedings Of The Grid Computing Environments Workshop (GCE). Https://Doi.Org/10.1109/GCE.2008.4738443

[39]. Foster, I., Zhao, Y., Raicu, I., & Lu, S. (2008). Cloud Computing And Grid Computing 360-Degree Compared. Proceedings Of The Grid Computing Environments Workshop (GCE). Https://Doi.Org/10.1109/GCE.2008.4738445

[40]. Stanoevska-Slabeva, K., & Wozniak, T. (2010). Cloud Basics — An Introduction To Cloud Computing. In K. Stanoevska-Slabeva, T. Wozniak, & S. Ristol (Eds.), Grid And Cloud Computing: A Business Perspective On Technology And Applications. Springer. Https://Doi.Org/10.1007/978-3-642-05193-7_2